Hidden Markov Models Selection Criteria based on Mean Field-like approximations

Florence Forbes, Nathalie Peyrard

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Hidden Markov Models Selection Criteria based on Mean Field-like approximations

Florence Forbes — Nathalie Peyrard

N° 4371

Février 2001

THÈME 4

ISSN 0249-6399 ISRN INRIA/RR--4371--FR+ENG

INRIA
RHÔNE-ALPES

Hidden Markov Models Selection Criteria based on Mean Field-like approximations

Florence Forbes , Nathalie Peyrard

Thème 4 — Simulation et optimisation
de systèmes complexes
Projet IS2

Rapport de recherche n° 4371 — Février 2001 — 33 pages

Abstract: Hidden Markov random fields appear naturally in problems such as image segmentation where an unknown class assignment has to be estimated from the observations for each pixel. Choosing the probabilistic model that best accounts for the observed data is an important first step for the quality of the subsequent estimation and analysis. A commonly used selection criterion is the Bayesian Information Criterion (BIC) of Schwarz (1978) but for hidden Markov random fields, its exact computation is not tractable due to the dependence structure induced by the Markov model. We propose approximations of BIC based on the mean field principle of statistical physics. The mean field theory provides approximations of Markov random fields by systems of independent variables leading to tractable computations. Using this principle, we first derive a class of criteria by approximating the Markov distribution in the usual BIC expression as a penalized likelihood. We then rewrite BIC in terms of normalizing constants (partition functions) instead of Markov distributions, which enables us to use finer mean field approximations and derive other criteria using optimal lower bounds for the normalizing constants. To illustrate the performance of our partition function-based approximation of BIC as a model selection criterion, we focus on the preliminary issue of choosing the number of classes before the segmentation task. Experiments on simulated and real data point out our criterion as promising: it takes spatial information into account through the Markov model and improves the results obtained with BIC for independent mixture models.

Florence Forbes is Researcher, Projet IS2, INRIA Rhône-Alpes, ZIRST, 655 av. de l'Europe, 38330 Montbonnot Saint-Martin, France. Email: Florence.Forbes@inrialpes.fr. Nathalie Peyrard is currently a member of the VISTA research team, doing a post-doc at IRISA, Campus de Beaulieu, 35042 Rennes, France. Email: npeyrard@irisa.fr.

Unité de recherche INRIA Rhône-Alpes
655, avenue de l'Europe, 38330 Montbonnot-St-Martin (France)
Téléphone : +33 4 76 61 52 00 — Télécopie +33 4 76 61 52 52

Key-words: Image segmentation, Hidden Markov random fields, Model selection, Bayesian Information Criterion, Mean field approximation, Partition function.

Critères de sélection de modèles de Markov cachés basés sur des approximations de type champ moyen.

Résumé : Les modèles de champs de Markov cachés apparaissent naturellement dans des problèmes tels que la segmentation d'image où il s'agit d'attribuer chaque pixel à une classe à partir des observations. Pour cela, choisir le modèle probabiliste qui prend le mieux en compte les données observées est primordial. Un critère de sélection de modèle communément utilisé est le Bayesian Information Criterion (BIC) de Schwarz (1978) mais dans le cas des champs de Markov cachés, la structure de dependance dans le modèle rend le calcul exact du critère impossible. Nous proposons des approximations de BIC qui se fondent sur le principe d'approximation en champ moyen issu de la physique statistique. La théorie du champ moyen fournit une approximation des champs de Markov par des systèmes de variables indépendantes pour lesquels les calculs sont alors faisables. À l'aide de ce principe, nous introduisons d'abord une famille de critères obtenus en approximant la loi markovienne qui apparaît dans l'expression usuelle de BIC sous forme de vraisemblance pénalisée. Nous considérons ensuite une réécriture de BIC en termes de constantes de normalisation (fonctions de partition) qui a l'avantage de permettre l'utilisation d'approximations plus fines. Nous en déduisons de nouveaux critères en utilisant des bornes optimales des fonctions de partitions. Pour illustrer les performances de ces derniers, nous considérons le problème du choix du bon nombre de classes pour la segmentation. Les résultats observés sur des données simulées et réelles sont prometteurs. Ils confirment que ce type de critères prend bien en compte l'information spatiale. En particulier, les résultats obtenus sont meilleurs qu'avec le critére BIC calculé pour des modèles de mélanges indépendants.

Mots-clés : Segmentation d'image, Champs de Markov cachés, Sélection de modèles, Critère d'information bayésien (BIC), Approximation du champ moyen, Fonction de partition.

Contents

List of Tables

List of Figures

1 Introduction

Problems involving incomplete data, where part of the data is missing or unobservable, are common in image analysis. The aim may be to recover an original image which is hidden and has to be estimated from a noisy or blurred version. More generally, the observed and hidden data are not necessarily of the same nature. The observations may represent measurements, *e.g.* multidimensional variables recorded for each pixel of an image while the hidden data could consist of an unknown class assignment to be estimated from the observations for each pixel. This case is usually referred to as image segmentation. In the context of statistical image segmentation, choosing the probabilistic model that best accounts for the observed data is an important first step for the quality of the subsequent estimation and analysis. In most cases the choice is done subjectively using expert knowledge or *ad hoc* procedures and there is a striking lack of systematic data-based approaches. We recast this choice as a problem of probabilistic model comparison and use the standard approach of Bayes factors. Evaluating the Bayes factor of one model against another involves calculating the ratio of the integrated likelihoods for each model, *i.e.* the likelihoods of the data integrated over the respective model parameters. For a lot of models of interest, these integrated likelihoods are high dimensional and intractable integrals so that most available software is generally inefficient for their evaluation. Various approximations have been proposed. In particular the Bayesian Information Criterion (BIC) approximation of Schwarz (1978) is based on the Laplace method for integrals. It leads to an equation giving the log-integrated likelihood as the maximized log-likelihood minus a correction (or penalization) term and an $O(1)$ error (as the sample size tends to infinity). BIC can be compared to other selection criteria. One of them is AIC (Akaike Information Criterion of Akaike 1973) which differs from BIC in the correction term but has been shown to overestimate the number of parameters in practice. The criterion proposed in Rissanen (1989) is based on stochastic complexity and is similar to BIC, and methods using cross validation (Zhang 1993) seem promising but their tractability in our context is not straightforward due to the dependence structure in the data. Many other approaches can be found in the literature on model selection (see for instance the list of references in Kass and Raftery 1995).

BIC has become quite popular due to its simplicity and its good results in cases where p-values and the standard model selection procedures based on them were unsatisfactory. In BIC, the $O(1)$ error does suggest the approximation to be somewhat crude. However empirical experience has found the approximation to be more accurate in practice than the

$O(1)$ error term would suggest. As regards model selection, Kass and Raftery (1995) observe that the criterion does not seem to be grossly misleading in a qualitative sense as long as the number of degrees of freedom involved in the comparison is relatively small relative to sample size. In this paper, we consider Markov model-based image segmentation and focuse on the use of BIC for the underlying issue of choosing a model from a collection of hidden Markov random fields. In this case, we have no specific results on the quality of BIC as an approximation of the integrated likelihood and this choice as a selection criterion is arguable. However, the question of the criterion ability to asymptotically choose the correct model can be addressed independently of the integrated likelihood approximation issue. As an illustration, Gassiat (2001) proved recently that for the more special but related case of hidden Markov chains, under reasonable conditions, the *maximum penalized marginal likelihood* estimator of the number of hidden states in the chain was consistent. This estimator is defined for a class of penalization terms that includes the BIC correction term and involves an approximation of the maximized log-likelihood which is not necessarily good, namely the maximized log-marginal likelihood. In particular, this criterion is consistent even if there is no guarantee that it provides a good approximation of the integrated likelihood. The choice of BIC for hidden Markov model selection appears then reasonable and we will show that criteria with good experimental behavior can be derived from it.

The difficulty in the context of hidden Markov random fields lies in that the maximized log-likelihood part in BIC involves Markov distributions whose exact computation requires an exponential amount of time. As regards observed Markov random fields selection, Ji and Seymour (1996) propose a consistent procedure based on penalized Besag pseudo-likelihood (Besag 1975) and mention few other previous works (see the references therein). When the fields are hidden, little has been done to address the selection problem. Two approximations of BIC are proposed in Stanford and Raftery (2001): for the PLIC (Pseudo-likelihood Information Criterion) the required maximized distribution is approximated by the Qian and Titterington pseudo-likelihood (Qian and Titterington 1991), while a simpler approximation, MMIC (Marginal Mixture Information Criterion) is based on the marginal distribution of pixel values. In practice, good results are reported for PLIC in Stanford and Raftery (2001) whereas MMIC is less satisfactory. In this paper, we propose approximations of BIC based on the mean field principle. Mean field theory of statistical physics (Chandler 1987) is an approach providing an approximation of a Markov random field by a system of independent variables and leading to tractable computations. We use a generalization of the mean field principle presented in a previous work (Celeux, Forbes, and Peyrard 2002) and derive a class

of criteria that includes PLIC as a particular case and as a result gives some new insight on its nature. We also show that the straightforward use of the mean field approximation can be improved by rewritting BIC in terms of normalizing constants (partition functions) instead of Markov distributions and then using optimal mean field lower bounds (Gibbs-Bogoliubov-Feynman bounds) for the normalizing constants. We derive this way an other tractable criterion denoted by BIC^{GBF}. Questions of interest relevant to model selection include choosing the Markov field neighborhood or more generally its energy function and choosing the number of classes in which to segment the data. They can all be addressed straightforwardly in our framework but we focus on the latter because of its practical importance. Experiments on simulated and real data point out BIC^{GBF} as a promising criterion. It is easy to compute and shows good and stable performance. It takes spatial information into account through the Markov model and improves the results obtained with BIC for independent mixture models. In particular, it seems to avoid the overestimation of the number of classes observed in Biernacki, Celeux, and Govaert (2000).

The complete parametric models for the observed and unobserved data are specified in Section 2 and the basics for BIC are recalled in Section 3. The mean field approximation principle is briefly presented in Section 4 and in Section 5 we show how we propose to use it to compute approximations of BIC and derive new computationaly tractable criteria for hidden Markov model selection. Experiments are reported in Section 6 and a discussion section ends the paper.

2 Hidden Markov models

Let S be a finite set of sites with a neighborhood system defined on it. Let $|S| = n$ denote the number of sites. A typical example in image analysis is the two dimensional lattice with a second order neighborhood system. For each site, the neighbors are the eight sites surrounding it. A set of sites C is called a clique if the sites are all neighbors. Let V be a finite set with K elements. Each of them will be represented by a binary vector of length K with one component being 1, all others being 0, so that V will be seen as included in $\{0,1\}^K$. We define a discrete Markov random field as a collection of discrete random variables, $\mathbf{Z} = \{Z_i,\ i \in S\}$, defined on S, each Z_i taking values in V, whose joint probability distribution satisfies the following properties,

$$\forall \mathbf{z}, \quad P_G(z_i \mid \mathbf{z}_{S\setminus\{i\}}) = P_G(z_i \mid z_j,\ j \in N(i)), \tag{1}$$

$$\forall \mathbf{z}, \quad P_G(\mathbf{z}) > 0, \tag{2}$$

where $\mathbf{z}_{S\backslash\{i\}}$ denotes a realization of the field restricted to $S\backslash\{i\} = \{j \in S, j \neq i\}$ and $N(i)$ denotes the set of neighbors of i. More generally, if A is a subset of S, we will write $\mathbf{z}_A$ for $\{z_i,\ i \in A\}$. In words, property (1) means that interactions between site i and the other sites actually reduce to interactions with its neighbors. Property (2) is important for the Hammersley-Clifford theorem to hold. This theorem states that the joint probability distribution of a Markov field is a Gibbs distribution (for which we use the notation P_G) given by

$$P_G(\mathbf{z}) \quad = \quad W^{-1} \exp(-H(\mathbf{z})), \tag{3}$$

where H is the energy function

$$H(\mathbf{z}) \quad = \quad \sum_c V_c(\mathbf{z}_c). \tag{4}$$

The sum is over the set of cliques and the V_c's are the clique potentials which may depend on parameters, not specified in the notation, $W = \sum_{\mathbf{z}} \exp(-H(\mathbf{z}))$ is the normalizing factor also called the partition function. We will write $\sum_{\mathbf{z}}$ (resp. $\sum_{\mathbf{z}_A}$) a sum over all possible values of $\mathbf{z}$ (resp. $\mathbf{z}_A$). The computation of W involves all possible realizations $\mathbf{z}$ of the Markov field. Therefore, it is in general exponentially complex and not computationally feasible. This can be a problem when using these models in situations where an expression of the joint distribution $P_G(\mathbf{z})$ is required. An approximation of the distribution (3) is the pseudo-likelihood introduced by Besag (1975) and defined as

$$\mathcal{PL}(\mathbf{z}) \quad = \quad \prod_{i \in S} P_G(z_i \mid \mathbf{z}_{N(i)}). \tag{5}$$

Each term in the product is easy to compute,

$$P_G(z_i \mid \mathbf{z}_{N(i)}) \quad = \quad \frac{\exp(-\sum_{c \ni i} V_c(\mathbf{z}_c))}{\sum_{z_i'} \exp(-\sum_{c \ni i} V_c(\mathbf{z}_c'))}, \tag{6}$$

with $\mathbf{z}_c' = \{z_i', z_j, j \in c, j \neq i\}$. Expression (5) is a genuine probability distribution only when the variables are independent but it can be used to obtain estimates of a Markov random field parameters. It has been used by Stanford (1999) in the model selection context (see Section 5). In Section 5, we will use other approximations based on systems of independent variables. Their factorization properties simplify computations as (5) and they correspond to valid probability models.

Image segmentation involves observed data and unobserved data to be recovered. The unobserved data is modeled as a discrete Markov random field, $\mathbf{Z}$, as defined in (3) with energy function H depending on a parameter β. In hidden Markov models, the observations $\mathbf{Y}$ are conditionally independent given $\mathbf{Z}$, according to a density f which is assumed to be of the following type (θ is a parameter and the f_i's are given),

$$\begin{aligned} f(\mathbf{y} \mid \mathbf{z},\ \theta) &= \prod_{i \in S} f_i(y_i \mid z_i,\ \theta) \\ &= \exp\{\sum_{i \in S} \log f_i(y_i \mid z_i,\ \theta)\}, \end{aligned} \tag{7}$$

assuming that all the $f_i(y_i \mid z_i,\ \theta)$ are positive. This makes the model similar to an independent mixture model (*cf.* McLachlan and Peel 2000). An independent mixture model could be seen as a hidden Markov model where the hidden field $\mathbf{Z}$ is one of independent identically distributed variables. In the general case, the complete likelihood is given by

$$\begin{aligned} P_G(\mathbf{y}, \mathbf{z} \mid \theta, \beta) &= f(\mathbf{y} \mid \mathbf{z}, \theta)\ P_G(\mathbf{z} \mid \beta) \\ &= W(\beta)^{-1} \prod_{i \in S} f_i(y_i \mid z_i,\ \theta) \prod_c \exp\{-V_c(\mathbf{z}_c \mid \beta)\} \\ &= W(\beta)^{-1} \exp\{-H(\mathbf{z} \mid \beta) + \sum_{i \in S} \log f_i(y_i \mid z_i,\ \theta)\}. \end{aligned} \tag{8}$$

Thus the conditional field $\mathbf{Z}$ given $\mathbf{Y} = \mathbf{y}$ is a Markov field as $\mathbf{Z}$ is. Its energy function is

$$H(\mathbf{z} \mid \mathbf{y}, \theta, \beta) = H(\mathbf{z} \mid \beta) - \sum_{i \in S} \log f_i(y_i \mid z_i,\ \theta). \tag{9}$$

In the following developments, we will refer to Markov fields $\mathbf{Z}$ and $\mathbf{Z}$ given $\mathbf{Y} = \mathbf{y}$ as the marginal and conditional fields and denote by $\Psi = (\theta, \beta)$ the vector parameter.

3 Bayesian Information Criterion

In a Bayesian framework, a way of selecting a model among R models $M_1, M_2, \ldots, M_R$ consists of choosing the model with highest posterior probability. By Bayes theorem, the posterior probability of M_r ($r \in \{1, \ldots, R\}$) given the data $\mathbf{y}$ is

$$P(M_r|\mathbf{y}) = \frac{P_G(\mathbf{y}|M_r)P(M_r)}{\sum_{k=1}^{R} P_G(\mathbf{y}|M_k)P(M_k)},$$

where $P_G(\mathbf{y}|M_r)$ is the integrated or marginal likelihood of model M_r and $P(M_r)$ is its prior probability. Assuming that all models have equal prior probabilities, choosing the model with

the highest posterior probability is equivalent to select the model with the largest integrated likelihood,

$$P_G(\mathbf{y}|M_r) = \int P_G(\mathbf{y}|\Psi_r, M_r)P(\Psi_r|M_r)d\Psi_r, \tag{10}$$

where Ψ_r varies in the model M_r parameter space and $P(\Psi_r|M_r)$ is the prior distribution on Ψ_r for the same model. Computing (10) is not usually tractable. A simple and often reliable way to approximate the integrated likelihood is provided by the Bayesian Information Criterion (BIC) of Schwarz (1978) (see for instance Kass and Raftery 1995),

$$2\log P_G(\mathbf{y}|M_r) \approx \mathrm{BIC}(M_r) = 2\log P_G(\mathbf{y} \mid \Psi_r^{ml}) - d_r \log(n), \tag{11}$$

where Ψ_r^{ml} is the maximum likelihood estimate of Ψ_r,

$$\Psi_r^{ml} \quad = \quad \arg\max_{\Psi_r} P_G(\mathbf{y} \mid \Psi_r, M_r),$$

and d_r is the number of free parameters in model M_r. It has been widely used in the context of selecting the number of components in independent mixture models (Fraley and Raftery 1998, Roeder and Wasserman 1997). In this context BIC limitations have been pointed out and in particular, it has been observed that in practice the criterion can tend to overestimate the right number of components when the true model is not in $\{M_1, \ldots . M_R\}$ (see Biernacki, Celeux, and Govaert 2000).

For Hidden Markov models the difficulty comes from that Ψ_r^{ml} and $P_G(\mathbf{y} \mid \Psi_r^{ml})$ are not available. For BIC, methods using simulations have been investigated in Newton and Raftery (1994) while Stanford (1999) proposed using the pseudo-likelihood (5) as an approximation to the intractable Markov distribution. In this paper, we suggest using the mean field approximation principle to derive a class of other tractable criteria. As for the pseudo-likelihood approximation, it consists of replacing the original Markov distribution by a product easier to deal with. We recall the mean field principle in the next section and describe applications in the model selection context in Section 5.

4 Mean field Theory

4.1 Mean field approximation principle

The mean field approximation (*e.g.* Chandler 1987) is originally a method of approximation for the computation of the mean of a Markov random field. It can be used to provide an

approximation of the distribution of a Markov random field. The idea when considering a particular site i is to neglect the fluctuations of the sites interacting with i. The resulting system behaves as one composed of independent variables for which computation gets tractable. More specifically, for all j different from i, the Z_j's are fixed to their mean value $I\!E_G(Z_j)$. However, these mean values are unknown and it is actually the goal of the approximation to compute them. Therefore, the method depends on a self-consistency condition which is that the mean computed based on the approximation must be equal to the mean used to define this approximation. The mean field approximation $P^{mf}(\mathbf{z})$ of the Gibbs distribution is then defined by replacing the exact mean values by the mean values in the approximation, denoted by $\bar{\mathbf{z}} = \{\bar{z}_j, j \in S\}$,

$$P^{mf}(\mathbf{z}) \;=\; \prod_{i \in S} P_i^{mf}(z_i), \tag{12}$$

where $P_i^{mf}(z_i) \;=\; P_G(z_i \mid \bar{\mathbf{z}}_{N(i)})$.

Applying the self-consistency condition leads to

$$\bar{\mathbf{z}} \;=\; g(\bar{\mathbf{z}}) \;=\; \left\{ \begin{array}{l} g_1(\{\bar{z}_j, j \in N(1)\}) \\ \vdots \\ g_n(\{\bar{z}_j, j \in N(n)\}) \end{array} \right. , \tag{13}$$

with $\forall i \in S,\ g_i(\{\bar{z}_j, j \in N(i)\}) \;=\; \sum_{z_i} z_i P_i^{mf}(z_i)$.

The mean field approximation consists of solving this fixed point equation and taking the solution $\bar{\mathbf{z}} = \{\bar{z}_i, i \in S\}$ as an estimate of the exact mean field. More details are given in Chandler (1987). In particular, it can be shown that P^{mf} minimizes the Kullback-Leibler divergence $I\!E_P[\log(\frac{P(\mathbf{Z})}{P_G(\mathbf{Z})})]$ over the set of probability distributions P that factorize. The mean field approximation can also be presented using the concept of *thermodynamic perturbation theory* and a variational principle. This approach presents the mean field approximation above as a zeroth-order approximation and provides a natural first order approximation for the partition function. We specify this first order in the following section. In Section 5.1, we derive BIC approximations based on $P_G \sim P^{mf}$ while in Section 5.2, we use the first order approximation of the partition function.

4.2 First order Mean field approximation of the partition function

Let H^{mf}, W^{mf} and $I\!E^{mf}$ denote respectively the energy function, the partition function and the expectation under model (12). An approach consists of performing a perturbation

theory with the mean field model (12) as the reference model or zeroth order model. In this approach, we assume that the fluctuations about the mean field energy H^{mf} are small, *i.e.* the difference $\Delta H = H - H^{mf}$ is small. The starting point is the following exact factorization of the partition function,

$$\begin{aligned} W &= \sum_z \exp(-(H^{mf}(z) + \Delta H(z))) \\ &= W^{mf} I\!E^{mf}[\exp(-\Delta H(Z))] \, . \end{aligned} \tag{14}$$

Since ΔH is small, the second factor in (14) can be expanded,

$$\begin{aligned} I\!E^{mf}[\exp(-\Delta H(Z))] &= I\!E^{mf}[1 - \Delta H(Z) + \ldots] \\ &= 1 - I\!E^{mf}[\Delta H(Z)] + \ldots \end{aligned}$$

The first terms of the right-hand side above are the first terms in the expansion around zero of $\exp(-I\!E^{mf}[\Delta H(Z)])$ so that neglecting the second and higher orders terms in ΔH leads to the following first order pertubation theory result. It gives an approximation of the exact partition function W,

$$W \approx W^{GBF} = W^{mf} \exp(-I\!E^{mf}[H(Z) - H^{mf}(Z)]) \, . \tag{15}$$

The quality of this approximation can be investigated through the following inequality, also called the Gibbs-Bogoliubov-Feynman (GBF) bound (Chandler 1987),

$$W \geq W^{mf} \exp(-I\!E^{mf}[H(Z) - H^{mf}(Z)]) \, . \tag{16}$$

Therefore, we always have $W \geq W^{GBF}$. As a first order approximation, we can expect W^{GBF} to be a closer approximation than W^{mf} which corresponds to the zeroth order. This is illustrated by the example in the Appendix where we compare the three quantities for a 2-color Potts model. Note that the same inequality is valid for any energy other than H^{mf}. However, the mean field model (12) is optimal among models with factorization property, in the sense that it maximizes the GBF bound in (16) for such models.

5 Mean field like approximations of BIC

The mean field approach consists of neglecting fluctuations from the mean in the environment of each pixel. More generally, we talk about mean field-like approximations when the value at site i does not depend on the values at other sites which are all set to constants (not

necessarily the means) independently of the value at site i (Celeux, Forbes, and Peyrard 2002). In Section 5.1, we apply this idea to release the computational burden when dealing with the intractable distribution $P_G(\mathbf{y} \mid \Psi)$ in BIC computation. This approach is the most straightforward considering expression (11) of BIC and includes criterion PLIC introduced in Stanford (1999) and recalled below. However, we will show that in practice, this does not always lead to satisfying results. In Section 5.2, we show that approximating the whole distribution is actually not necessary and we derive alternative criteria approximating BIC using the first order partition function approximation (15). Experimental results confirm the superiority of this method.

As regards the notation, we consider a model M_r among R hidden Markov models ($r = 1, \ldots, R$) as defined by (3) and (7) with parameters $\Psi_r = (\theta_r, \beta_r)$.

5.1 Approximating the Gibbs distribution

A mean field like approximation of a Gibbs distribution can be defined as follows. Given a configuration $\tilde{\mathbf{z}}$, set the neighbors to $\tilde{\mathbf{z}}_{N(i)}$ and replace the marginal distribution $P_G(\mathbf{z} \mid \beta_r)$ by

$$P_{\tilde{\mathbf{z}}}(\mathbf{z} \mid \beta_r) = \prod_{i \in S} P_G(z_i \mid \tilde{\mathbf{z}}_{N(i)}, \beta_r). \tag{17}$$

It corresponds to an observed likelihood of the form

$$\begin{aligned} P_{\tilde{\mathbf{z}}}(\mathbf{y} \mid \Psi_r) &= \sum_{\mathbf{z}} f(\mathbf{y} \mid \mathbf{z}, \theta_r) P_{\tilde{\mathbf{z}}}(\mathbf{z} \mid \beta_r) \\ &= \prod_{i \in S} \sum_{z_i} f_i(y_i \mid z_i, \theta_r)\ P_G(z_i \mid \tilde{\mathbf{z}}_{N(i)}, \beta_r) \\ &= \prod_{i \in S} P_G(y_i \mid \tilde{\mathbf{z}}_{N(i)}, \Psi_r). \end{aligned} \tag{18}$$

We consider $P_{\tilde{\mathbf{z}}}(\mathbf{y} \mid \Psi_r)$ as a candidate for an approximation of the intractable $P_G(\mathbf{y} \mid \Psi_r)$ involved in expression (11) of BIC. The flexibility of our proposition is then in the choice of the values $\tilde{\mathbf{z}}$. A natural candidate would be one that leads to a reasonable approximation of $P_G(\mathbf{y}, \mathbf{z} \mid \Psi_r)$. In our model, $P_G(\mathbf{z} \mid \beta_r)$ and $P_G(\mathbf{z} \mid \mathbf{y}, \Psi_r)$ are not available while $f(\mathbf{y} \mid \mathbf{z}, \theta_r)$ is. Knowing $f(\mathbf{y} \mid \mathbf{z}, \theta_r)$, it is enough to approximate one of the unknown quantities, either $P_G(\mathbf{z} \mid \beta_r)$ or $P_G(\mathbf{z} \mid \mathbf{y}, \Psi_r)$, to derive an approximation of the other and of the joint distribution. Therefore, our selection of $\tilde{\mathbf{z}}$ can be driven by the quality of the corresponding approximation of $P_G(\mathbf{z} \mid \beta_r)$ or $P_G(\mathbf{z} \mid \mathbf{y}, \Psi_r)$. As regards the Kullback-Leibler divergence, the approximations cannot be both optimal and satisfy the Bayes rule. It seems

more reasonable to base our choice on the conditional field distribution rather than on the marginal field distribution. It has the advantage of taking the observations directly into account. Moreover, the study of the case of the homogeneous isotropic Potts model gives reasons disuading from using the mean field approximation on the marginal field (see Archer and Titterington 2000 and Celeux, Forbes, and Peyrard 2002).

For computing BIC, it then remains the problem of computing the maximum likelihood estimator Ψ_r^{ml}. Let $\hat{\Psi}_r$ be an approximation of Ψ_r^{ml}. An approximation for BIC is then,

$$\mathrm{BIC}^{\tilde{\mathbf{z}}}(\hat{\Psi}_r) = 2\log P_{\tilde{\mathbf{z}}}(\mathbf{y} \mid \hat{\Psi}_r) - d_r \log(n). \tag{19}$$

As regards the quality of such an approximation, it is not clear whether $\tilde{\mathbf{z}}$ and $\hat{\Psi}_r$ must be chosen independently or not. As an example, the Pseudo-Likelihood Information Criterion (PLIC) of Stanford (1999) is a particular case of $\mathrm{BIC}^{\tilde{\mathbf{z}}}(\hat{\Psi}_r)$. Indeed, if the unsupervised ICM algorithm is used to get an estimate Ψ_r^{ICM} and a restoration $\mathbf{z}_r^{ICM}$ and then $\hat{\Psi}_r$ and $\tilde{\mathbf{z}}$ set to these values, approximation (19) becomes

$$\begin{aligned} \mathrm{BIC}^{\mathbf{z}_r^{ICM}}(\Psi_r^{ICM}) &= 2\log(P_{\mathbf{z}^{ICM}}(\mathbf{y} \mid \Psi_r^{ICM})) - d_r \log(n) \\ &= \mathrm{PLIC}(M_r). \end{aligned} \tag{20}$$

In this case $\hat{\Psi}_r$ and $\tilde{\mathbf{z}}$ were computed using a single iterative procedure (ICM) which alternates between estimating Ψ_r and estimating $\mathbf{z}$ so that the final estimates can be deduced from one another. In this paper, we propose to use for $\hat{\Psi}_r$ the output of the EM algorithm-based procedures described in Celeux, Forbes, and Peyrard (2002) and referred to as *mean field like algorithms* in what follows. Like ICM, the algorithms alternatively produce a configuration $\tilde{\mathbf{z}}$ and using (18) an estimation $\hat{\Psi}_r$. We then found natural to set $\tilde{\mathbf{z}}$ to the ones used in our procedures, corresponding either to an approximation of the conditional mean, the conditional mode or to a simulated realization of the conditional distribution. Based on the study in Celeux, Forbes, and Peyrard (2002), we chose simulated $\tilde{\mathbf{z}}$ through the *simulated field* algorithm which showed good performance as regards hidden Markov random fields parameter estimation and outperformed ICM in this task in most cases.

PLIC shows promising results when used to select the number of components in tests on synthetic and real images reported in Stanford (1999). In Section 6, we report additional results for $\tilde{\mathbf{z}}$ and $\hat{\Psi}_r$ set to values provided by the ICM algorithm (PLIC). The results when $\tilde{\mathbf{z}}$ and $\hat{\Psi}_r$ are obtained via mean field-like algorithms are not reported but can be found in Peyrard (2001). They were satisfying for real data but surprisingly unstable, as regards the number of components, on simulated data (simulated Potts models as in Section 6.1).

In this first approach, the use of the simulated field algorithm for $\hat{\Psi}_r$ appears reasonable and we will keep this estimation procedure in the next section. As regards the quality of the approximation of $P_G(\mathbf{y} \mid \Psi_r)$ by $P_{\tilde{\mathbf{z}}}(\mathbf{y} \mid \Psi_r)$, it is not easy to assess but in what follows we will propose a more satisfying alternative.

5.2 Approximating the partition function

In this section, the idea is to use an expression for BIC that involves only partition functions so that the problem of approximating the Markov distributions can be replaced by that of approximating the partition functions. The advantage is that the partition function first order approximations presented in Section 4.2 can be used and results in better approximations.

Let $W(\mathbf{y}, \Psi)$ and $W(\beta)$ be the partition functions for the conditional and marginal fields respectively,

$$\begin{aligned} W(\beta) &= \sum_{\mathbf{z}} \exp(-H(\mathbf{z}|\beta)) \\ W(\mathbf{y}, \Psi) &= \sum_{\mathbf{z}} \exp(-H(\mathbf{z}|\mathbf{y}, \Psi)). \end{aligned}$$

Using notation of Section 2, it comes from

$$P_G(\mathbf{y} \mid \Psi) = \frac{P_G(\mathbf{y}, \mathbf{z} \mid \Psi)}{P_G(\mathbf{z} \mid \mathbf{y}, \Psi)} = \frac{f(\mathbf{y} \mid \mathbf{z}, \theta)\ P_G(\mathbf{z} \mid \beta)}{P_G(\mathbf{z} \mid \mathbf{y}, \Psi)}$$

that

$$P_G(\mathbf{y} \mid \Psi) = \frac{f(\mathbf{y} \mid \mathbf{z}, \theta)\ \exp(-H(\mathbf{z}|\beta))}{\exp(-H(\mathbf{z}|\mathbf{y}, \Psi))}\ \frac{W(\mathbf{y}, \Psi)}{W(\beta)},$$

which using (9) simplifies into

$$P_G(\mathbf{y} \mid \Psi) = \frac{W(\mathbf{y}, \Psi)}{W(\beta)}. \tag{21}$$

Therefore expression (11) of BIC is equivalent to the following one which uses only the partition functions $W(\mathbf{y}, \Psi)$ and $W(\beta)$,

$$\mathrm{BIC}(M_r) = 2 \log W(\mathbf{y}, \Psi_r^{ml}) - 2 \log W(\beta_r^{ml}) - d_r \log(n). \tag{22}$$

Let $H^{mf}(\mathbf{z}|\beta)$ and $H^{mf}(\mathbf{z}|\mathbf{y}, \Psi)$ denote the mean field expressions for the marginal and conditional field energies. Using the first order approximations for the partition functions, a

new approximation of BIC is:

$$\begin{aligned} \text{BIC}^{GBF}(\hat{\Psi}_r) &= 2\log W^{mf}(\mathbf{y}, \hat{\Psi}_r) - 2I\!E^{mf}[H(\mathbf{Z}|\mathbf{y}, \hat{\Psi}_r) - H^{mf}(\mathbf{Z}|\mathbf{y}, \hat{\Psi}_r)|\mathbf{y}] \\ &\quad -2\log W^{mf}(\hat{\beta}_r) + 2I\!E^{mf}[H(\mathbf{Z}|\hat{\beta}_r) - H^{mf}(\mathbf{Z}|\hat{\beta}_r)] \\ &\quad -d_r \log(n). \end{aligned} \tag{23}$$

As before, $\hat{\Psi}_r$ must be estimated and we used mean field-like algorithms and more specifically, the simulated field algorithm (Celeux, Forbes, and Peyrard 2002). The marginal and conditional mean field approximations were then computed, using this value of the parameter, by solving the corresponding fixed point equations (13).

The expression of BIC^{GBF} (23) is more satisfying than approximation $\text{BIC}^{\tilde{\mathbf{Z}}}$ (19). A way to see the improvement is to rewrite $P_{\tilde{\mathbf{Z}}}(\mathbf{y} \mid \Psi_r)$ in (18) using partition functions as in (21),

$$P_{\tilde{\mathbf{Z}}}(\mathbf{y} \mid \Psi_r) = \frac{W_{\tilde{\mathbf{Z}}}(\mathbf{y}, \Psi_r)}{W_{\tilde{\mathbf{Z}}}(\beta_r)}.$$

Expressions for both quantities in the ratio are easily deduced from (17) and (6). The ratio (21) is thus better approximated in BIC^{GBF} than in $\text{BIC}^{\tilde{\mathbf{Z}}}$ since as explained in Section 4.2, it uses the best lower bound (16) for each partition function. Therefore, there are some theoretical and experimental reasons to believe that our BIC^{GBF} is a better approximation of the true BIC than PLIC. BIC^{GBF} is based on a better approximation of $P_G(\mathbf{y} \mid \Psi_r)$ and the procedure it uses to compute $\hat{\Psi}_r$ has shown to be as reliable if not better than ICM in practice (Celeux, Forbes, and Peyrard 2002). However, note that as regards model selection, this does not necessarily ensure that the resulting criterion would lead to better results although the experiments reported in Section 6 tend to confirm this.

6 Experiments

In this section, the gain in approximating the partition functions (BIC^{GBF}-like criteria) rather than the whole Markov distribution (PLIC-like criteria) is investigated. We examine the performance of the two approaches as regards the problem of choosing the number of classes in the segmentation. We report experiments on three types of images. For all examples, the observed images are considered as realizations of the simple following hidden Markov model. The distribution of the hidden field is supposed to be a K-color Potts model where each z_i takes one of K states, which represent K different class assignments or colors. Recall that each of the states is represented by a binary vector of length K with one

component being 1, all others being 0. The distribution of a K-color Potts model is defined by,

$$P_G(\mathbf{z} \mid \beta) \;=\; W(\beta)^{-1} \exp(\beta \sum_{i \sim j} z_i^t z_j) \;, \tag{24}$$

where β is a real non-negative parameter and the notation $i \sim j$ represents all couples of sites (i, j) which are neighbors.

For the f_i's we considered Gaussian distributions. If site i is in class k, f_i is the Gaussian distribution with mean μ_k and standard deviation σ_k. The parameter to be estimated is then $\{\beta, \theta\}$ with $\theta = \{(\mu_k, \sigma_k), k = 1, \ldots, K\}$. Let M_K be the model defined above when the number of colors is K. To assess its ability to select a relevant number K, the criterion is computed for model M_K with $K = K_{min}$ to $K = K_{max}$. The required estimations of $\hat{\Psi}_K$ for each value of K considered were obtained with the simulated field algorithm and $\mathrm{BIC}^{GBF}(\hat{\Psi}_K)$ was computed as defined in (23). We also report values of BIC when the images are seen as realizations of independent mixture models in order to measure the gain of taking spatial information into account when selecting the number of classes. The EM algorithm was used to estimate the parameters and the criterion (computed exactly in this case) is denoted by BIC^{IND}. We also compared with PLIC based on the ICM algorithm as an alternative criterion assuming a spatial model.

When not otherwise specified, the algorithms (Simulated field, EM and ICM algorithms) were initialized using the same segmentation computed by simple thresholding. We divided the pixel values range, in the degraded image, into regular intervals and assigned each of them to a component. The algorithms were all stopped after $N = 100$ iterations.

The images used for the experiments are described below. In Section 6.1, we first compare the criteria on fully simulated data. The models used for the simulations are the models used for the segmentations. In Section 6.2, we consider synthetic images degraded with some simulated Gaussian noise. The true K is known but the images are not realizations of a known probabilistic model. In Section 6.3 real-life images are considered.

6.1 Hidden K-color Potts models

We first tested the criteria on images simulated from hidden Potts models for which the true parameters β and θ were known. We created 100×100 images by simulating (Gibbs Sampling of Geman and Geman 1984) 2D K-color Potts models (24) for $K = 2, \ldots 6$ and different values of β, and then adding a Gaussian noise. We chose β so that the simulated

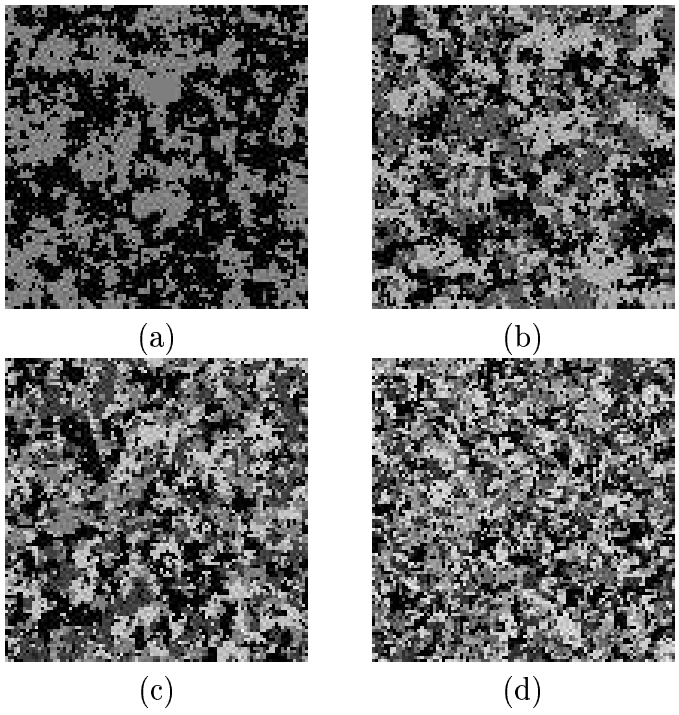

Figure 1: Simulations of a K-color Potts model for different values of K and β : (a) $K = 2$, $\beta = 0.78$, (b) $K = 3$, $\beta = 0.9$, (c) $K = 4$, $\beta = 1$, (d) $K = 5$, $\beta = 1$.

images present homogeneous regions and some spatial structure (*e.g.* Figure 1) for in other cases we cannot really expect the criteria to recover the true K. For smaller values of β typical realizations look much noisier and are visually close to independently distributed colors. For larger values, the simulations lead to close to monocolor images whatever the true K used for the simulations. We considered a first order neighborhood, *i.e.* four neighbors for each pixel. The simulated data correspond to hidden K-color Potts models for which $\theta = \{(\mu_k, \sigma_k), k = 1, \dots K\}$ with $\mu_k = k$ and $\sigma_k = 0.5$, for $k = 1, \dots K$. We used our knowledge of a constant variance, for the K states, to fit a model and recover the true image. For each model considered, 100 simulations were carried out. The corresponding criteria results are reported in Table 1. It appears that criteria BIC^{GBF} and PLIC perform well and outperform BIC^{IND} which shows degradation in selecting the right number of colors when K is larger than 4. This confirms the advantage of using spatial models even through approximations but does not enable to differentiate BIC^{GBF} from PLIC. More differences appear in the next two sections.

$K=2$, $\beta=0.78$	
selected K	2
BIC^{IND}	100
PLIC	100
BIC^{GBF}	100

$K=3$, $\beta=0.9$	
selected K	3
BIC^{IND}	100
PLIC	100
BIC^{GBF}	100

$K=4$, $\beta=1$			
selected K	3	4	5
BIC^{IND}	38	62	0
PLIC	0	100	0
BIC^{GBF}	0	99	1

$K=5$, $\beta=1$			
selected K	4	5	6
BIC^{IND}	79	21	0
PLIC	0	100	0
BIC^{GBF}	0	92	8

$K=6$, $\beta=1.1$				
selected K	4	5	6	7
BIC^{IND}	13	80	7	0
PLIC	0	2	98	0
BIC^{GBF}	0	0	99	1

Table 1: Degraded K-color Potts model: Selected K using BIC for independent mixture models (BIC^{IND}), pseudo-likelihood (PLIC) and mean field-like (BIC^{GBF}) approximations of BIC. The reported values are the number of times a given K is selected out of 100 experiments.

6.2 Noise-corrupted synthetic images

In this section, we consider noise-corrupted images corresponding to known values of K. Image (b) of Figure 2 is a 128×128 image obtained by adding some Gaussian noise to the 4-color image (a) of Figure 2. The noise parameters are given by $\theta = \{(\mu_k, \sigma_k), k = 1, \ldots, 4\}$ with $\mu_k = k$ and $\sigma_k = 0.5$ for $k = 1, \ldots, 4$. The other example (Figure 3) is a 133×142 noise-corrupted 2-color image. We used Gaussian densities with class-dependent variances so that the true noise parameters are $(\mu_1, \sigma_1) = (51, 130)$ and $(\mu_2, \sigma_2) = (255, 300)$. These images before degradation are not realizations from a known Markov field model. For estimation, we assumed a model with second order neighborhood (*i.e.* the eight closest neighbors for each pixel). The selected K are reported in Table 2. In these experiments, BIC^{GBF} and PLIC behave differently. We observe that BIC^{GBF} is better in selecting the right number of colors for images presenting thin features (*e.g.* Figure 3) while they both perform well when images are made of larger regions (*e.g.* Figure 2). Additional experiments were carried out with other images containing thin lines and showed similar results in favor of BIC^{GBF}.

6.3 Grey-level images

We eventually tried the criteria on real images for which it does not exist a true value for K (in real-life, it is usually part of the problem to assess its value) but for which intuition or expert knowledge could give an indication of what would be a reasonable value. As an

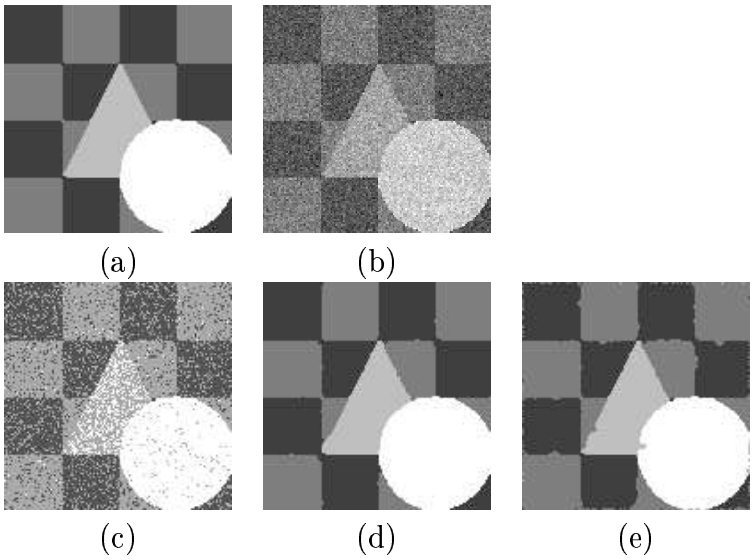

Figure 2: Checkerboard image : (a) original image, (b) noise-corrupted image, (c) 3-color segmentation using EM for independent mixtures, (d) and (e) 4-color segmentations using the simulated field and ICM algorithms.

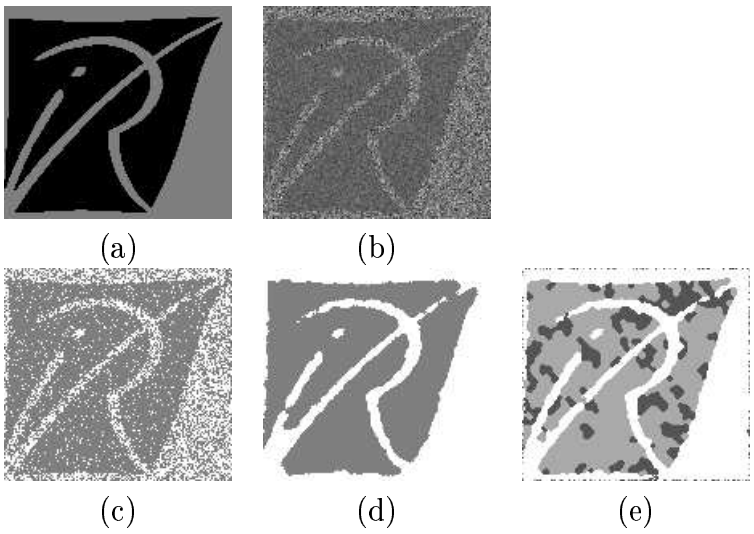

Figure 3: Logo image: (a) original image, (b) noise-corrupted image, (c) and (d) 2-color segmentations using EM for independent mixtures and the simulated field algorithm, (e) 3-color segmentation using ICM.

Checkerboard image	
criterion	selected K
BIC^{IND}	3
PLIC	4
BIC^{GBF}	4

Logo image	
criterion	selected K
BIC^{IND}	2
PLIC	3
BIC^{GBF}	2

Table 2: Noise-corrupted synthetic images: Selected K using BIC for independent mixture models (BIC^{IND}), pseudo-likelihood (PLIC) and mean field-like (BIC^{GBF}) approximations of BIC.

illustration, Figure 4 (a) is an aerial 100×100 image of a buoy against a background dark water and Figure 5 (a) is a 128×128 PET image of a dog lung (see Stanford (1999) for more details on their nature and origin).

For the first image, we suspect that 2 is a relevant value for K. Image 4 (a) presents some artifact (horizontal scan lines from the imaging process). Some pre-processing step to remove this known artifact could be carried out as in Stanford (1999) but we tested here the criteria on the raw data. The selected K are shown in Table 3 and the corresponding segmentations in Figure 4. BIC^{GBF} performs much better than PLIC which selects a too large number of components while BIC^{IND} probably suffers from not taking into account the spatial information, as can be seen on segmentation (d). These results were obtained using basic thresholding to produce initial segmentations for the estimation algorithms (simulated field and ICM algorithms). We tried BIC^{GBF} and PLIC with more refined initializations using the independent mixtures EM algorithm segmentations as first images. This can be seen as a pre-processing step. The selected K was then 2 for BIC^{GBF} (Figure 4 (c)) but still too large (7) for PLIC which leads to a meaningless segmentation (Figure 4 (f)).

For the dog lung image, the aim is to distinguish the lung from the rest of the image in order to measure the heterogeneity of the tissue in the region of interest. Only pixels in this delimited area will then be considered to compute a heterogeneity measure, such as a coefficient of variation. PLIC and BIC^{GBF} select rather different K with again a too large value for PLIC (Table 3). The corresponding segmentations are shown in Figure 5 (b) and (f). The 3-color segmentation obtained using BIC^{GBF} and the simulated field algorithm is the more satisfying as regards interpretation. It shows one color for the background and two for the lung itself. This is not surprising since the image is constructed based on radioactive emissions from gas in the lung. The two segments account for the high gas density in the interior of the lung and the somewhat lower gas density around the periphery. BIC^{IND}

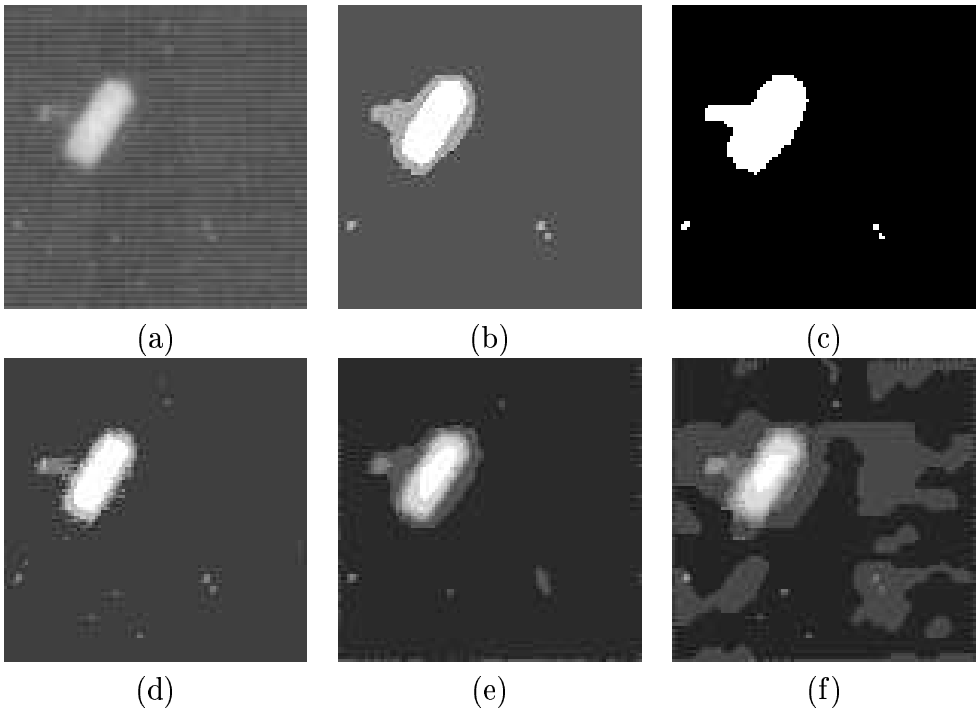

Figure 4: Buoy image : (a) original image, (b) and (c) 3 and 2-color segmentations using the simulated field algorithm respectively initialized by thresholding and EM for independent mixtures, (d) 4-color segmentation using EM for independent mixtures, (e) and (f) 6 and 7-color segmentations using ICM respectively initialized by thresholding and EM for independent mixtures.

Buoy image	
criterion	selected K
BIC^{IND}	4
PLIC	6
BIC^{GBF}	3

Dog lung image	
criterion	selected K
BIC^{IND}	3
PLIC	6
BIC^{GBF}	3

Table 3: Grey-level images: Selected K using BIC for independent mixture models (BIC^{IND}), pseudo-likelihood (PLIC) and mean field-like (BIC^{GBF}) approximations of BIC.

also selects 3 colors but the corresponding segmentation is rather different focusing more on the artificial background circle. We then also computed BIC^{GBF} and PLIC using the independent mixtures EM segmentations instead of the ones obtained via thresholding as initializing images but noticed no significant difference.

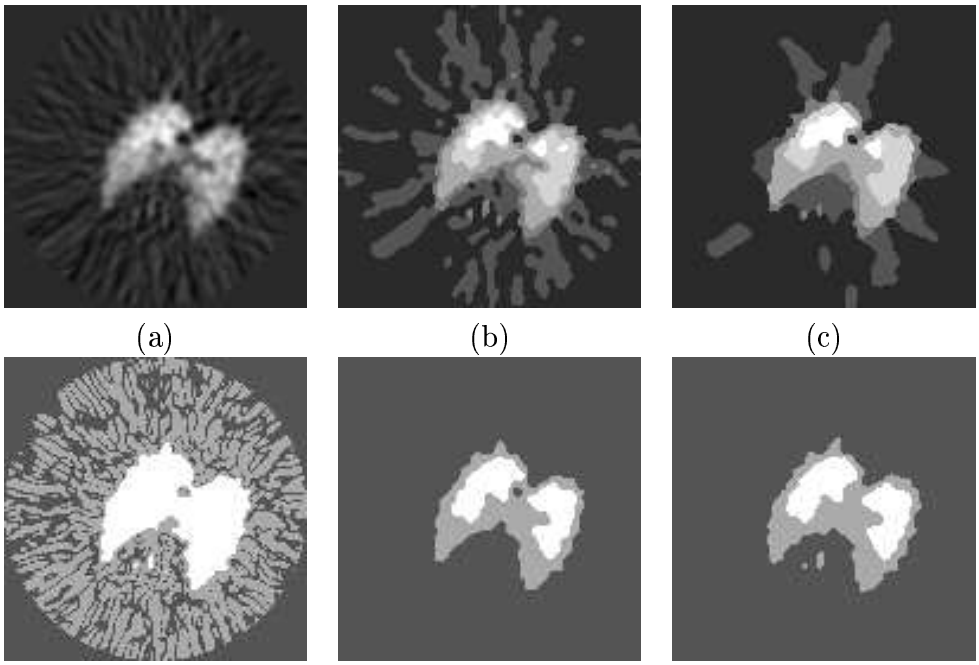

Figure 5: PET Image of a dog lung: (a) original image, (d) 3-color segmentation using EM for independent mixtures, (b) and (e) 6-color and 3-color segmentations using ICM, (c) and (f) 6-color and 3-color segmentations using the simulated field algorithm.

7 Discussion

In the context of Markov model selection, starting from BIC as our selection criterion, we proposed using mean field-like approximations to deal with the computation of the intractable Markov distribution in BIC expression. More specifically, one of our contributions was to notice that BIC could be rewritten in terms of partition functions for which a first order rather than zeroth order mean field approximation was available (Section 5.2). The advantage is that the quality of the approximation is easier to assess since it uses the best lower bounds for the partition functions. We introduced a class of new criteria among which we chose one, the so-called BIC^{GBF} (equation (23)) based on these theoretical considerations regarding the quality of the approximation of the intractable likelihood and based on previous experimental results as regards parameter estimation for a various types of images. First it appears that taking spatial information into account leads to some improvements when compared to BIC for independent mixture models (BIC^{IND}). Then, our criterion differentiates from PLIC (BIC approximation based on the pseudo-likelihood) in its ability to deal better with thin features in images. It also shows good performance on real images although we can suspect decreasing performance in the presence of artifact (like scan lines) that the criterion may consider as relevant information instead of noise. However this is likely to be handle by some pre-processing step using reasonable initializations (EM for independent mixtures).

After carrying out various experiments, it appeared that a sensible procedure for model selection would be to first perform simple procedures. For example, for selecting the number of components into which to segment an image, a natural procedure is the EM algorithm for independent mixtures models easy to implement and for which BIC values can be computed exactly. In some cases, this could lead to reasonably satisfying selection and segmentation so that users may choose not to go further. If not, as it is likely to occur for images with significative spatial structure, the corresponding procedure could possibly be further used to initialize more refined algorithms based on spatial models. For example, Stanford and Raftery (2001) studied ICM and used the pseudo-likelihood approximation while we propose to use the simulated field algorithm of Celeux, Forbes, and Peyrard (2002) and the mean field approximation principle to compute criterion BIC^{GBF}. On the set of images tested in our experiments, our procedure showed much better performance especially on real data. We believe that this is mainly due to a better approximation of the likelihood in BIC^{GBF} (see the Appendix for an illustration of the superiority of the first order approximation) coupled to a satisfying estimation of the parameter provided by the simulated field algorithm.

This study remains somewhat limited in that it is mainly exploratory and based on experiments. We did not address the question of the consistency of the various criteria. As far as we know no such results are currently available for hidden Markov random fields. In some recent work, Gassiat (2001) consider a maximized penalized marginal likelihood criterion for estimating the number of hidden states in hidden Markov chains. Gassiat (2001) proves a consistency result for this criterion although the marginal likelihood involved is not necessarily close to the likelihood (they are equal only when the variables are independent). This suggests that a good approximation of the maximized log-likelihood is not a strong requirement to obtain consistent criteria. A key point in Gassiat (2001) seems to be the decomposition of the criterion as a sum of identically distributed terms. The criteria proposed in this paper can also be written as sum because of the factorization property of the distributions involved. The generalization is not straightforward but our next step is therefore to investigate if consistency results can be deduced in a similar way.

Acknowledgements

The authors are grateful to G. Celeux for many valuable comments.

References

Akaike, M. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrox and F. Caski (Eds.), *Second International Symposium on Information Theory*, pp. 267.

Archer, G. E. B. and D. M. Titterington (2000). Parameter estimation for hidden Markov chains. *To appear in Journal of Statistical Planning Inference*.

Besag, J. (1975). Statistical analysis of non-lattice data. *The Statistician 24*, 179–195.

Biernacki, C., G. Celeux, and G. Govaert (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence 22*, 719–725.

Celeux, G., F. Forbes, and N. Peyrard (2002). EM procedures using mean field-like approximations for Markov model-based image segmentation. *To appear in Pattern Recognition*.

Chandler, D. (1987). *Introduction to Modern Statistical Mechanics*. Oxford University Press.

Fraley, C. and A. Raftery (1998). How many clusters? Which clustering method? Answers via mo del-based cluster analysis. *Computer Journal 41*, 578–588.

Gassiat, E. (2001). Likelihood ratio inequalities with applications to various mixtures. Technical Report 2001-20, Mathematiques,Orsay.

Geman, S. and D. Geman (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence 6*, 721–741.

Ji, C. and L. Seymour (1996). A consistent model selection procedure for Markov random fields based on penalized pseudolikelihood. *Annals of Applied Probability 6*, 423–443.

Kass, R. and A. Raftery (1995). Bayes factor. *Journal of the American Statisticial Association 90*, 733–795.

McLachlan, G. J. and D. Peel (2000). *Finite Mixture Models*. Wiley.

Newton, M. and A. Raftery (1994). Approximate Bayesian Inference by the Weighted Likelihood Bootstrap (with discussion). *Journal of the Royal Statistical Society, Serie B 56*, 3–48.

Peyrard, N. (2001). *Approximations de type champ moyen des modèles de champ de Markov pour la segmentation de données spatiales*. Ph. D. thesis, U.F.R. d'informatique et de mathématiques appliquées, Université Joseph Fourier, Grenoble I, France.

Qian, W. and D. M. Titterington (1991). Estimation of parameters in hidden Markov models. *Phil. Trans. R. Soc. Lond. A* (337), 407–428.

Rissanen, J. (1989). Stochastic complexity in statistical inquiry. *Word Scientific, Teaneck, New Jersey*.

Roeder, K. and L. A. Wasserman (1997). Practical Bayesian Density Estimation Using Mixtures of Normals. *Journal of the American Statistical Association 92*, 894–902.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics 6*, 461–464.

Stanford, D. (1999). *Fast Automatic Unsupervised Image Segmentation and Curve Detection in Spatial Point Processes*. Ph. D. thesis, Department of Statistics, University of Washington, Seattle.

Stanford, D. and A. Raftery (February 2001). Determining the Number of Colors or Gray Levels in an Image Using Approximate Bayes Factors: The Pseudolikelihood Information Criterion (PLIC). Technical report, Department of Statistics, University of Washington, http://www.stat.washington.edu/.

Zhang, P. (1993). Model selection via multifold cross validation. *The Annals of Statistics 21*, 299–313.

Appendix: Zeroth and first order approximations for the partition function of a 2-color Potts model

The notation is that of Section 4.2. Considering simple Potts models, our aim is to illustrate that W^{GBF} (equation (15)) can be a better approximation of W than the standard mean field approximation W^{mf}. The energy of a Potts model can be written

$$H(\mathbf{z}|\beta) = -\beta \sum_{i\sim j} z_i^t z_j = -\frac{\beta}{2} \sum_{i=1}^{n} z_i^t \sum_{j\in N(i)} z_j \ ,$$

it follows the zeroth order mean field approximation

$$H^{mf}(\mathbf{z}|\beta) = -\beta \sum_{i=1}^{n} z_i^t \sum_{j\in N(i)} \overline{z}_j,$$

with $\overline{z}_j = I\!E^{mf}[Z_j]$. Then

$$I\!E^{mf}[H(\mathbf{Z}|\beta)] = -\frac{\beta}{2} \sum_{i=1}^{n} \overline{z}_i^t \sum_{j\in N(i)} \overline{z}_j = \frac{1}{2} I\!E^{mf}[H^{mf}(\mathbf{Z}|\beta)],$$

so that

$$W^{mf} \ = \ \sum_{\mathbf{z}} \exp(-H^{mf}(\mathbf{z}|\beta)) \ = \ \sum_{i=1}^{n} \sum_{z_i} \exp(\beta z_i^t \sum_{j\in N(i)} \overline{z}_j),$$

and

$$W^{GBF} \ = \ W^{mf} \exp(I\!E^{mf}[H(\mathbf{Z}|\beta)]) \ = \ W^{mf} \exp(-\frac{\beta}{2} \sum_{i=1}^{n} \overline{z}_i^t \sum_{j\in N(i)} \overline{z}_j).$$

Using symmetries, for all $i = 1, \ldots, n$, we can write $\overline{z}_i = \mathbf{m}$ with $\mathbf{m}$ being, in the two-color case, the two-component vector $(m_1, m_2)^t$ satisfying $m_1 + m_2 = 1$ and the following consistency conditions,

$$\begin{aligned} m_1 &= \frac{\exp(\beta N m_1)}{\exp(\beta N m_1) + \exp(\beta N m_2)} \\ m_2 &= \frac{\exp(\beta N m_2)}{\exp(\beta N m_1) + \exp(\beta N m_2)} , \end{aligned}$$

where $N = |N(i)|$ is the number of neighbors assumed the same for all sites.
This is equivalent to solve

$$\begin{aligned} m_1 &= \frac{\exp(\beta N m_1)}{\exp(\beta N m_1) + \exp(\beta N (1 - m_1))} \\ &= \frac{1}{1 + \exp(\beta N (1 - 2m_1))} . \end{aligned} \tag{25}$$

Note that if m_1 satisfies (25), then $1 - m_1$ is also solution. For $\beta < K/N$, *i.e.* $\beta < 2/N$ there is only one solution $m_1 = 1/2$. For $\beta > 2/N$ there are two additional solutions m_1 and $1 - m_1$ with $m_1 > 1/2$. We focus on solutions $m_1 \neq 1/2$. Such a solution is a non-constant function of β whose closed form expression is not available. However, using (25), β can be expressed as a function f of m_1 given by

$$\beta = f(m_1) = \frac{1}{N(1 - 2m_1)} \log(\frac{1 - m_1}{m_1}). \tag{26}$$

It is easy to check that $f(1 - m_1) = f(m_1)$ so that two symmetric solutions lead to the same β as expected. We can also check that $f(m_1)$ tends to $2/N$ when m_1 tends to $1/2$ and to infinity when m_1 tends to 1. The graph of m_1 *wrt* β is shown in Figure 6.

The quantity m_1 appears in the expressions of W^{mf} and W^{GBF}, while the true W depends only on β. However, when W is available in closed form, using (26), the three quantities can be expressed and compared as functions of m_1. For periodic boundary conditions, it comes

$$W^{mf} = (\exp(\beta N m_1) + \exp(\beta N (1 - m_1)))^n \tag{27}$$

and

$$W^{GBF} = W^{mf} \exp(-\frac{\beta}{2} N n (m_1^2 + (1 - m_1)^2)) . \tag{28}$$

It follows, using (25)

$$\begin{aligned} \log(W^{mf}) &= \beta N n m_1 + n \log(1 + \exp(\beta N (1 - 2m_1))) \\ &= \beta N n m_1 - n \log(m_1) \end{aligned} \tag{29}$$

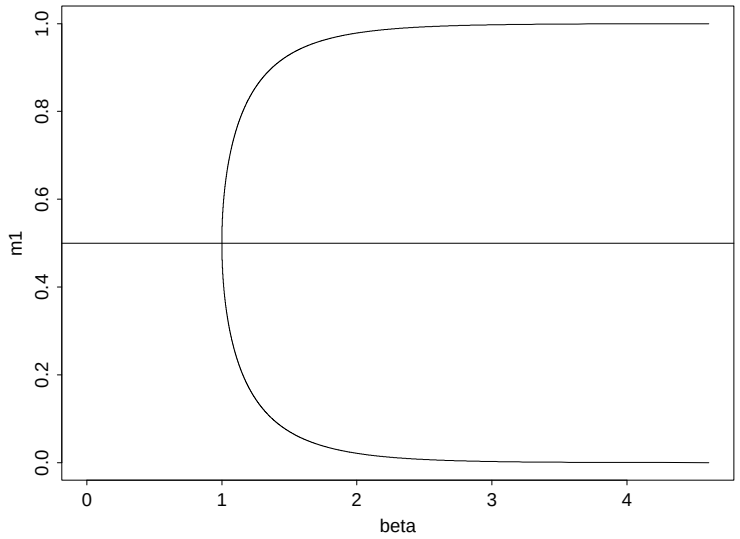

Figure 6: 1-dimensional 2-color Potts model: solutions $(m_1, 1-m_1)$ of the mean field consistency conditions as β varies.

and

$$\log(W^{GBF}) = \frac{\beta}{2}Nn(4m_1 - 2m_1^2 - 1) + n\log(1+\exp(\beta N(1-2m_1))) \ . \quad (30)$$

As regards W, a closed form is not available in general. However, in the 1-dimensional case for which $N = 2$ an expression of W is

$$W = (\exp(\beta)+1)^n + (\exp(\beta)-1)^n \ .$$

It is then easy to compare the logarithms. For $N = 2$,

$$\begin{aligned}
\log(W^{mf}) &= 2nm_1\beta + n\log(1+\exp(2\beta(1-2m_1))) \\
\log(W^{GBF}) &= n(4m_1 - 2m_1^2 - 1)\beta + n\log(1+\exp(2\beta(1-2m_1))) \\
\log(W) &= n\beta + n\log(1+\exp(-\beta)) + \log\left(1+\left(\frac{1-\exp(-\beta)}{1+\exp(-\beta)}\right)^n\right) \ .
\end{aligned}$$

For $\beta < 1$, $m_1 = 1/2$, it comes

$$\begin{aligned}
\log(W^{mf}) &= n\beta + n\log(2) \ , \\
\log(W^{GBF}) &= n\frac{\beta}{2} + n\log(2) \ .
\end{aligned}$$

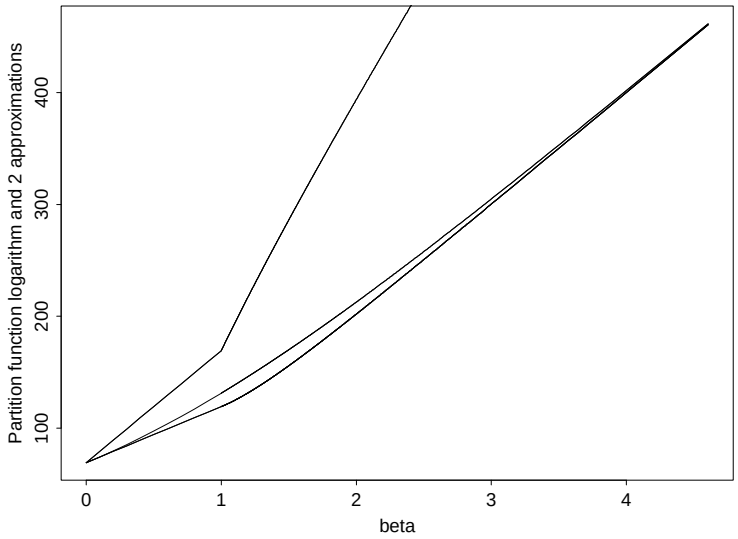

Figure 7: 1-dimensional ($N = 2$) 2-color Potts model with $n = 100$ sites: exact partition function logarithm (middle curve) and two approximations for $\beta > 0$. The closer curve corresponds to $\log W^{GBF}$ and the other one to $\log W^{mf}$.

The corresponding graphs are shown in Figure 7.

When $\beta > 1$ there are no analytical expression for m_1 as a function of β but we can plot the graphs by inverting (26) (See Figure 7). Note that $\log(W^{mf})$ and $\log(W^{GBF})$ remain the same when m_1 is changed to $1 - m_1$. It appears clearly on the plot that $\log(W^{GBF})$ is a far better approximation of the exact $\log(W)$ than $\log(W^{mf})$.

For dimension greater than 1, the mean field approximation expressions (27) and (28) are still valid but the computation of the true W is exponentially complex. We restricted then to a 3×3 grid, *i.e.* $n = 9$ sites and considered successively $N = 4$ and $N = 8$ neighbors. For $N = 4$, the exact partition function is,

$$W = 102\exp(6\beta) + 144\exp(8\beta) + 198\exp(10\beta) + 48\exp(12\beta) + 18\exp(14\beta) + 2\exp(18\beta) .$$

For $N = 8$, it comes

$$W = 252\exp(16\beta) + 168\exp(18\beta) + 72\exp(22\beta) + 18\exp(28\beta) + 2\exp(36\beta).$$

The partition function logarithm and its approximations are shown in Figures 8 and 9. In the general case, when β tends to infinity, W behaves (if K denotes the number of colors) as $K\exp(nN\beta/2)$, which is the dominant term in the sum over all possible configurations.

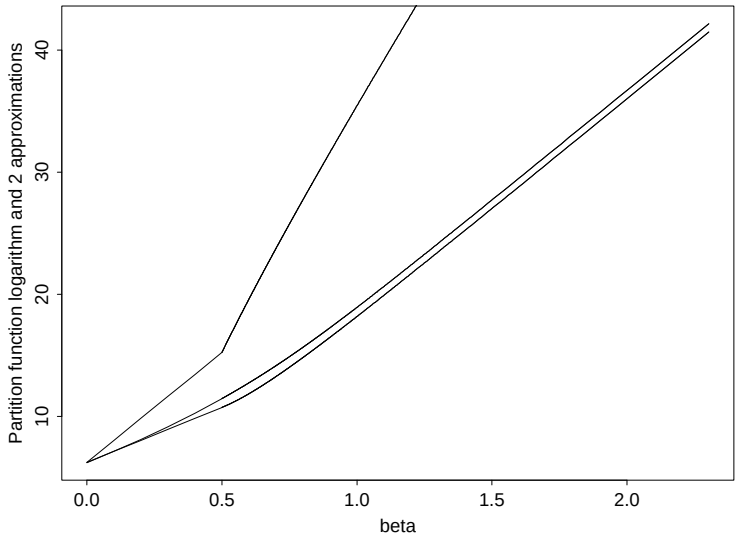

Figure 8: 2-color Potts model on a 3×3 grid, $N = 4$ neighbors: exact partition function logarithm (middle curve) and two approximations for $\beta > 0$. The closer curve corresponds to $\log W^{GBF}$ and the other one to $\log W^{mf}$.

The term $nN/2$ is the maximum number of homogeneous cliques. It occurs for each of the K monocolor configurations. Therefore, when β tends to infinity $\log(W)$ behaves as $nN\beta/2 + \log K$. When looking at expressions (29) and (30) it appears that when β tends to infinity, m_1 tends to 0 or 1 and in both cases $\log(W^{mf})$ behaves as $nN\beta$ and $\log(W^{GBF})$ as $nN\beta/2$. This suggests ways to improve the approximations. The $\log(2) = 0.69$ difference between $\log(W)$ and $\log(W^{GBF})$ appears more clearly on Figures 8 and 9. Again, $\log(W^{GBF})$ appears to be a much better approximation of $\log(W)$ than $\log(W^{mf})$.

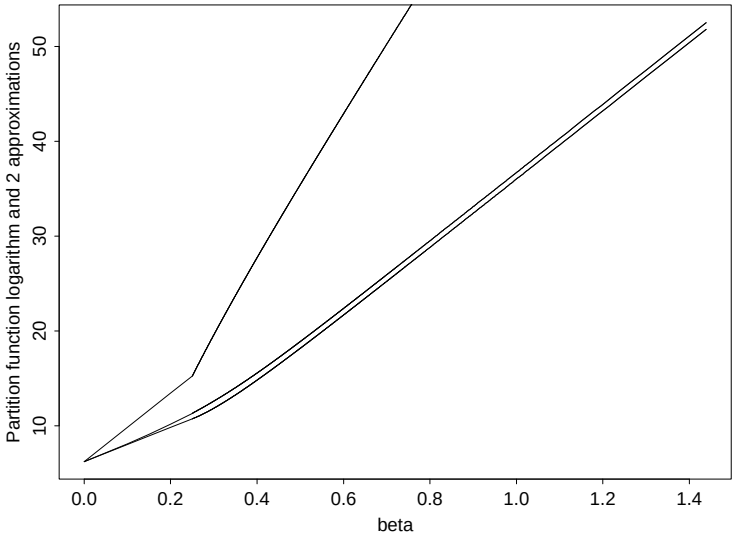

Figure 9: 2-color Potts model on a 3×3 grid, $N = 8$ neighbors: exact partition function logarithm (middle curve) and two approximations for $\beta > 0$. The closer curve corresponds to $\log W^{GBF}$ and the other one to $\log W^{mf}$.

Unité de recherche INRIA Rhône-Alpes
655, avenue de l'Europe - 38330 Montbonnot-St-Martin (France)

Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)
Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399